Group 2
Siva Satya Sri Ganesh Seeram,
Hansi Abeynanda,
Eren Berk Kama,
Irshad Ahmed Meer,
Zinat Behdad

EP3260: Machine Learning Over Networks
Homework Assignment 2
Due Date: February 15, 2023

Problem 2.1

Consider Human Activity Recognition Using Smartphones dataset $\{(\mathbf{x}_i, y_i)\}_{i\in[N]}$, with inputs defined as the accelerometer and gyroscope sensors, and outputs defined as moving (e.g., walking , running, dancing) or not (sitting or standing). Consider the logistic ridge regression loss function

$$\underset{\mathbf{w}}{\text{minimize}} \quad f(\mathbf{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2,$$

where $f_i(\mathbf{w}) = \log\left(1 + \exp\{-y_i\mathbf{w}^T\mathbf{x}_i\}\right)$.

Then, address the following questions:

(a) Is f Lipschitz continuous? If so, find a small B?

(b) Is f_i smooth? If so, find a small L for f_i? What about f?

(c) Is f strongly convex? If so, find a high μ?

(a) f is Lipschitz continuous if

Lipschitz continuity (bounded gradients)

$$\|\boldsymbol{w}\|_2 \le D \Rightarrow \|\nabla f(\boldsymbol{w})\|_2 \le B$$

I or $\|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2 \le D \Rightarrow |f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1)| \le B\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$

$$\nabla f(w) = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(w) + 2\lambda w$$

$$\nabla f_i(w) = \frac{-y_i x_i \, e^{-y_i w^T x_i}}{1+e^{-y_i w^T x_i}} = \frac{-y_i x_i}{1+e^{y_i w^T x_i}} \; < e$$

$$\|\nabla f_i(w)\| = \left\| \frac{-y_i x_i}{1+e^{y_i w^T x_i}} \right\| \le |y_i| \cdot \|x_i\|$$

$$\|\nabla f(w)\|_2 = \left\| \frac{1}{N} \sum_{i\in[N]} \nabla f_i(w) + 2\lambda w \right\|_2 \le B$$

$$\le \frac{1}{N} \sum_{i\in[N]} \underbrace{\|\nabla f_i(w)\|}_{\le |y_i| \cdot \|x_i\|} + 2\lambda \underbrace{\|w\|}_{\le D}$$

$$\le \underbrace{\frac{1}{N} \sum |y_i| \cdot \|x_i\| + 2\lambda D}_{B}$$

(b) $\|\nabla f_i(w_2) - \nabla f_i(w_1)\|_2 =$

$$= \left\| \frac{-y_i x_i e^{-y_i w_2^T x_i}}{1+e^{-y_i w_2^T x_i}} - \frac{-y_i x_i e^{-y_i w_1^T x_i}}{1+e^{-y_i w_1^T x_i}} \right\|$$

$$= \|y_i\| \cdot \|x_i\| \left\| \frac{1}{1+e^{y_i w_2^T x_i}} - \frac{1}{1+e^{y_i w_1^T x_i}} \right\| \quad (1)$$

Let $h(w) = \dfrac{1}{1+e^{y_i w^T x_i}}$

$h(w)$ is continuous and differentiable

$\therefore$ By mean value theorem, $\exists\, z \in (w_1, w_2)$ s.t

$$h(w_2) - h(w_1) = \nabla h(z)^T (w_2 - w_1)$$

$$\| h(w_2) - h(w_1) \| \le \| \nabla h(z)\| \cdot \| w_2 - w_1 \| \quad (2)$$

where $\nabla h(w) = \dfrac{-y_i x_i e^{y_i w^T x_i}}{(1+e^{y_i w^T x_i})^2}$

$\therefore$ (2) $\Rightarrow$

$$\| h(w_2) - h(w_1)\| \le \|y_i x_i\| \underbrace{\left\| \frac{e^{y_i z^T x_i}}{(1+e^{y_i z^T x_i})^2} \right\|}_{\le 1} \cdot \|w_2 - w_1\|$$

$$\le |y_i| \|x_i\| \|w_2 - w_1\|$$

$\therefore$ ① $\Longrightarrow$

$$\|\nabla f_i(\omega_2) - \nabla f_i(\omega_1)\| = |y_i| \, \|x_i\| \, \|h(\omega_2) - h(\omega_1)\|$$

$$\leq \underbrace{|y_i|^2 \cdot \|x_i\|^2}_{L} \, \|\omega_2 - \omega_1\|$$

$\therefore$ a small L for f_i is $|y_i|^2 \cdot \|x_i\|^2$

About f:

∇f is also Lipschitz continuous

Proof:

$$\|\nabla f(\omega_2) - \nabla f(\omega_1)\| =$$

$$\Big\| \frac{1}{N} \sum_{i \in [N]} \big(\nabla f_i(\omega_2) - \nabla f_i(\omega_1)\big) + 2\lambda(\omega_2 - \omega_1) \Big\|$$

$$\leq \frac{1}{N} \sum_{i \in [N]} \|\nabla f_i(\omega_2) - \nabla f_i(\omega_1)\| + 2\lambda \|\omega_2 - \omega_1\|$$

$$\leq \underbrace{\left[\frac{1}{N} \sum_{i \in [N]} |y_i|^2 \cdot \|x_i\|^2 + 2\lambda \right] \|\omega_2 - \omega_1\|}_{L_1}$$

$\Longrightarrow f$ is L_1-smooth

(C) f is strongly convex, because

$\lambda \|w\|^2$ is strongly convex and

$\frac{1}{N} \sum_{i \in [N]} f_i(w)$ is convex.

Let $g(w) = \lambda \|w\|^2 = \lambda w^T w$

$\nabla g(w) = 2\lambda w \implies \nabla^2 g(w) = 2\lambda I$

$\implies g$ is strongly convex, with $\boxed{\mu = 2\lambda}$

$\therefore f$ is also 2λ-strongly convex

Problem 2.2

Let us assume that there exist scalars $c_0 \geq c > 0$ such that for all $k \in \mathbb{N}$

$$\nabla f(\mathbf{w}_k)^T \mathbb{E}_{\zeta_k}\left[g(\mathbf{w}_k;\zeta_k)\right] \geq c\|\nabla f(\mathbf{w}_k)\|_2^2, \tag{1a}$$

$$\|\mathbb{E}_{\zeta_k}\left[g(\mathbf{w}_k;\zeta_k)\right]\|_2 \leq c_0\|\nabla f(\mathbf{w}_k)\|_2. \tag{1b}$$

Furthermore, let us assume that there exist scalars $M \geq 0$ and $M_V \geq 0$ such that for all $k \in \mathbb{N}$

$$\mathrm{Var}_{\zeta_k}\left[g(\mathbf{w}_k;\zeta_k)\right] \leq M + M_V\|\nabla f(\mathbf{w}_k)\|_2^2. \tag{2}$$

For the convergence proof of SGD with an L-smooth convex objective function (see slides), prove that

$$\mathbb{E}_{\zeta_k}\left[\|g(\mathbf{w}_k;\zeta_k)\|_2^2\right] \leq \alpha + \beta\|\nabla f(\mathbf{w}_k)\|_2^2.$$

We want to find a relationship between (1a), (1b) and (2) and give relation and prove that it holds.

Using the definition of variance:

$$\mathrm{Var}_{\zeta_k}\left[g(\omega_k;\zeta_k)\right] = \mathbb{E}\left\{\left[g(\omega_k;\zeta_k) - \mathbb{E}\{g(\omega_k;\zeta_k)\}\right]^2\right\}$$

$$= \mathbb{E}\{\|g(\omega_k;\zeta_k)\|_2^2 - \mathbb{E}\{g(\omega_k;\zeta_k)\}^T g(\omega_k;\zeta_k)$$
$$- g(\omega_k,\zeta_k)^T \mathbb{E}\{g(\omega_k;\zeta_k)\} + \|\mathbb{E}\{g(\omega_k,\zeta_k)\}\|^2\}$$
$$\leq M + M_V \|\nabla f(\omega_k)\|_2^2$$

Therefore :

$$\mathbb{E}\{\|g(w_k;\xi_k)\|_2^2\} \leq \underbrace{2\,\mathbb{E}\{\mathbb{E}\{g(w_k;\xi_k)\}^T g(w_k;\xi_k)\}}_{2\|\mathbb{E}\{g(w_k,\xi_k)\}\|_2^2}$$

$$\underbrace{-\|\mathbb{E}\{g(w_k,\xi_k)\}\|^2}+M+M_v\|\nabla f(w_k)\|_2^2$$

$$=\underbrace{\|\mathbb{E}\{g(w_k,\xi_k)\}\|_2^2}+M+M_v\|\nabla f(w_k)\|_2^2$$

using (1b) : $\left(\|\mathbb{E}\{g(w_k,\xi_k)\}\|_2\right)^2 \leq \left(c_0\|\nabla f(w_k)\|_2\right)^2$

taking square of (1b), we obtain

$$\mathbb{E}\{\|g(w_k,\xi_k)\|_2^2\} \leq M+(c_0^2+M_v)\|\nabla f(w_k)\|_2^2$$

Therefore, we have $\alpha = M$, $\beta = c_0^2 + M_v$ and the relation is satisfied

Problem 2.3

For the SGD with non-convex objective function, prove that with square summable but not summable step-size, we have for any $K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2\right] < \infty \tag{3}$$

and therefore

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]} a_k} \sum_{k\in[K]} \alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2\right] \xrightarrow{K\to\infty} 0 \tag{4}$$

3-) We want to prove the given relation for SGD with non-convex objective function and square summable but not summable step-size.

As step size is square summable, $\{\alpha_k\} \to 0$.

Therefore, we may take $\alpha_k L . M_G \leq \mu$. $\forall k \in \mathbb{N}$

We use the approach in references and look at,

$$E\{f(w_{k+1})\} - E\{f(w_k)\} \leq -(\mu - \frac{1}{2}\alpha_k L . M_G) \cdot \alpha_k E\{\|\nabla f(w_k)\|_2^2$$

$$+ \frac{1}{2}\alpha_k^2 L . M$$

$$\leq -\frac{1}{2}\mu\alpha_k E\{\|\nabla f(w_k)\|_2^2\}$$

$$+ \frac{1}{2}\alpha_k^2 L . M$$

As in p 3.22. of lecture notes, we sum the relations, and obtain,

$$f_{inf} - E\{f(w_1)\} \le E\{f(w_{K+1})\} - E\{f(w_1)\}$$
$$\le -\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k E\{\|\nabla f(w_k)\|_2^2\} + \frac{1}{2} L.M. \sum_{k=1}^{K} \alpha_k^2$$

Further using algebraic manipulations,

$$\sum_{k=1}^{K} \alpha_k E\{\|\nabla f(w_k)\|_2^2\} \le \frac{2(E\{f(w_1)\} - f_{inf})}{\mu} + \frac{LM}{\mu} \sum_{k=1}^{K} \alpha_k^2$$

as α_k are square summable, we have

$$\sum_{k=1}^{K} \alpha_k^2 < \infty .$$

It converges to a finite limit as K increases:

Which shows that the first relation holds.

As α_k is not summable,

$$\sum_{k=1}^{K} \alpha_k = \infty$$

Tends to infinity. Therefore,

$$E\left\{\left(\frac{1}{\sum_{k=1}^{K}\alpha_k}\right)\cdot\sum_{k=1}^{K}\alpha_k \|\nabla f(w_k)\|_2^2\right\} \xrightarrow{K\to\infty} 0$$

As the summation is in denominator.